A Probabilistic Model of Face Mapping with Local Transformations and Its Application to Person Recognition

Florent Perronnin, Jean-Luc Dugelay, *Senior Member*, *IEEE*, and Kenneth Rose, *Fellow*, *IEEE*

Abstract—This paper proposes a new measure of "distance" between faces. This measure involves the estimation of the set of possible transformations between face images of the same person. The global transformation, which is assumed to be too complex for direct modeling, is approximated by a patchwork of local transformations, under a constraint imposing consistency between neighboring local transformations. The proposed system of local transformations and neighboring constraints is embedded within the probabilistic framework of a two-dimensional hidden Markov model. More specifically, we model two types of intraclass variabilities involving variations in facial expressions and illumination, respectively. The performance of the resulting method is assessed on a large data set consisting of four face databases. In particular, it is shown to outperform a leading approach to face recognition, namely, the Bayesian intra/extrapersonal classifier.

Index Terms—Biometrics, face recognition, image processing, hidden Markov model, distance.

1 Introduction

Let us consider the general pattern classification problem where a sample x is to be assigned to one of a set of possible classes $\{\omega_i\}$. Within the Bayesian decision framework, the optimal classifier (commonly referred to as the minimum risk classifier) employs the decision rule: assign observed pattern x to the class ω_i that minimizes the conditional risk $R(\omega_i|x) = \sum_j \lambda(\omega_i|\omega_j)p(\omega_j|x)$, where the loss function $\lambda(\omega_i|\omega_j)$ quantifies the loss incurred for selecting ω_i when the true class of x is ω_j, and where $p(\omega_j|x)$ is the (posterior) probability of class ω_j given that sample x was observed, which is computed in practice from $p(\omega_j)$—the class prior probabilities—and $p(x|\omega_j)$—the class-conditional probability density functions (pdf). For a detailed review see, e.g., [1]. Typically, the loss functions are determined by the application and are hence assumed known, but the class priors and class-conditional pdf's need to be estimated given a training set of labeled samples. In practice, the more challenging task is the estimation of class-conditional pdf's that characterize intraclass variability, and its accuracy is a primary determining factor for the classifier performance. The quality of these estimates hinges on two crucial factors: the correctness of the chosen model and the availability of a sufficiently large training set to estimate the model parameters. Obviously, these two considerations are interrelated as the fewer parameters of a compact model will require less training data to be robustly estimated.

The discipline of biometrics is concerned with the automatic recognition of a person based on his/her physiological or behavioral characteristics [2]. Biometric applications involve pattern classification systems where the samples are biometric data from a person under consideration, and need to be classified into categories whose nature depends on the specific task at hand. For the *identification* task, a new biometric sample is assigned to the most likely identity from a predefined set of identities. In this case, the classes are the possible identities. For the *verification* task, the system is probed with a biometric sample and a claimed identity. The goal is to decide whether the sample indeed corresponds to the claimed identity. Verification is thus a two-class decision problem where the classes correspond to the acceptance/rejection decision.

The focus of this paper is on face recognition [3], [4], a central area in biometrics. It is a very challenging task, as faces of different persons share global shape characteristics, while face images of the same person are subject to considerable variability, which might overwhelm the measured interperson differences. Such variability is due to a long list of factors including facial expressions, illumination conditions, pose, presence or absence of eyeglasses and facial hair, occlusion, and aging. Although much progress has been made over the past three decades, face recognition is largely considered an open problem, as observed during the FERET evaluation [5] and the facial recognition vendor tests (FRVT) 2000 [6] and 2002 [7], and is a highly active research topic.

Data scarcity is often a problem of paramount importance in biometric applications. When a new user first enrolls in a system, only a few instances of the considered biometrics are typically captured in order to reduce the

- *F. Perronnin is with Xerox Research Centre Europe, Image Processing Group, 6 chemin de Maupertuis, 38240 Meylan, France. E-mail: Florent.Perronnin@xrce.xerox.com.*
- *J.-L. Dugelay is with the Multimedia Communications Department, Institut Eurecom, 2229 Route des Crêtes, BP 193, 06904 Sophia-Antipolis, France. E-mail: jean-luc.dugelay@eurecom.fr.*
- *K. Rose is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106-9560. E-mail: rose@ece.ucsb.edu.*

Manuscript received 13 Apr. 2004; revised 10 Nov. 2004; accepted 13 Jan. 2005; published online 12 May 2005.
Recommended for acceptance by R. Basri.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0174-0404.

duration of enrollment and minimize inconvenience to the user (as well as maximize user cooperation). Hence, very little intraclass variability can be observed during the enrollment session. If only one sample is provided, intraclass variability is obviously impossible to assess. In the case of face recognition, the image which is provided (or its representation) is thus directly used as a template and the likelihood $p(x|\omega_i)$ can be interpreted as a possible measure of similarity between the query and enrollment images. More generally, we note that the main issue is the ability to define a distance between images which is meaningful for the task at hand. While many algorithms focus on the problem of representation, i.e., feature extraction, less attention has been given to the derivation and computation of an appropriate distance. For instance, the popular Eigenfaces [8] and Fisherfaces algorithms [9], [10], [11] employ low dimensional coding of face images. The distance between faces in the face subspace is based on simple metrics such as L_1, L_2, cosine and Mahalanobis distances [12], [13]. Combinations thereof such as the $\text{Mahalanobis-}L_1$, $\text{-}L_2$ and -cosine [14] have been proposed. Variations, such as the "Yambor" distance [14] for Eigenfaces or the weighted euclidean distance for Fisherfaces [9], [15], have also been suggested. The candidate distance that yields the best results in a given set of experiments is simply chosen. However, it is often difficult to ascertain why one distance measure performs better than another.

To define a meaningful distance, it is beneficial to formalize the relationship between observations of the same class, i.e., between face images of the same person. Due to the scarcity of data, we have to assume (or postulate) the existence of a "universal" distance measure that can be applied to different classes, i.e., that the intraclass variability is similar in the various classes. Thus, the parameters of the distance measure can be estimated from a larger training set which is not restricted to images of persons that are enrolled in the system. If O_t denotes the template image for class ω_i, O_q a query image, and $\mathcal{R}$ the relationship between images of the same class, then the class-conditional probability is expressed as:

$$p(O_q|\omega_i) = p(O_q|O_t, \mathcal{R}). \quad (1)$$

A distance based on the above expression has already been used by the Bayesian intra/extrapersonal classifier [16], [17], [18] that aims at estimating the distribution of image differences and by related approaches such as [19]. Note that, while the Elastic Graph Matching (EGM) [20] also defines a distance between face images, it does not make use of a probabilistic framework. However, other approaches related to EGM, such as [21], have made an attempt to define a probabilistic distance.

In this paper, we consider a novel measure of "distance" between faces. This measure involves the estimation of the set of possible transformations between face images of the same person. The global transformation is assumed too complex for direct modeling and is approximated with a set of local transformations under a constraint imposing consistency between neighboring local transformations. The proposed local transformations and neighboring constraints are embedded within the probabilistic framework of a two-dimensional hidden Markov model (2D HMM). This general framework is specialized to two types of intraclass variability: facial expressions and illumination variations. In the proposed system, they are modeled separately using different types of local transformations.

The remainder of the paper is organized as follows: In the next section, we provide a more detailed description of the general framework of the proposed face recognition system. In Sections 3 and 4, respectively, we specialize this framework to the problems of face recognition in the presence of facial expressions and illumination variations. In Section 5, we relate this work to existing work in the face recognition literature. In Section 6, we provide experimental results involving 4 databases (the FERET [5], PIE [22], Yale B [23], and AR [24] face databases) and more than 10,000 images. Conclusions are drawn and presented in the last section.

2 Framework

Our premise is that a global transformation between two images may be too complex to be modeled directly and that it should be approximated with a set of *local transformations*. These local transformations should be as simple as possible for efficient implementation but the composition of all local transformations (i.e., the global transformation) should be rich enough to model a wide range of variabilities between face images of the same person. However, if we do not restrict the set of admissible combinations of local transformations, the model might become overflexible and "succeed" to patch together very different faces. This observation naturally leads to the second component of our framework: the *neighborhood coherence constraint* whose purpose is to provide context information and to impose consistency requirements on the combination of local transformations. It must be emphasized that such neighborhood consistency rules introduce dependencies in the local transformation selection for the various image regions, and the optimal solution must therefore involve a global decision. To combine the local transformation and consistency costs, we propose to embed the system within a probabilistic framework using 2D HMMs. Note that HMMs have already been successfully applied to the problems of face detection and face recognition [25], [26], [27], [28]. However, the approach we propose is fundamentally different as our focus is on modeling a transformation between face images while the goal of the previously cited approaches is to model the face.

Let us assume that feature vectors are extracted on a grid from the query image O_q. At any location on O_q, the system is assumed to be in some unknown state. If we assume that the 2D HMM is first-order Markovian, the state of the system at a given position depends on the state of the system at the adjacent positions in both horizontal and vertical directions, as quantified by the *transition probabilities*. At each position, an observation is emitted according to the state-conditional *emission probabilities*. In our framework, local transformations are identified with the states of the HMM, and emission probabilities model the local

mapping cost. These transformations are "hidden" and information on them can only be extracted from the observations. Transition probabilities relate states of neighboring regions and implement the consistency rules.

The set of possible global transformations and, hence, the resulting distance, primarily depends on the allowed local transformations. In this paper, we consider in particular two types of local transformations: *grid* and *feature* transformations. A grid transformation consists of a local deformation of the feature extraction lattice of the query image. A feature transformation consists of transforming the extracted features directly through the application of a meaningful operator. Note that, if we work in a transform domain, a feature transformation can reflect both geometric or photometric transformations in the pixel domain.

We will next specialize this framework for two very different types of variabilities: elastic facial distortions (such as expressions), using grid transformations, and illumination variations, using feature transformations.

3 Modeling Facial Expressions

Elastic distortions due to facial expressions will be modeled through grid transformations. In Sections 3.1 and 3.2, respectively, we consider the emission probabilities and the transition probabilities of our HMM. Finally, in Section 3.3, we briefly introduce the turbo hidden Markov model (T-HMM) as an efficient approximation of the computationally intractable 2D HMM. The T-HMM framework provides efficient approximate formulas to 1) compute the score $P(O_q|O_t,\mathcal{R})$ and 2) estimate the parameters of $\mathcal{R}$.

3.1 Emission Probabilities

Let $o_{i,j}$ be the observation extracted from O_q at position (i,j) on the grid with $1 \leq i \leq I$ and $1 \leq j \leq J$. Let $q_{i,j}$ be the associated state. It τ is a translation vector, the emission probability, i.e., the probability that at position (i,j) the system emits observations $o_{i,j}$ given that it is in state $q_{i,j} = \tau$, is denoted $b^{\tau}_{i,j} = P(o_{i,j}|q_{i,j} = \tau, \mathcal{R})$. If we examine our score $P(O_q|O_t,\mathcal{R})$, it is clear that the HMM parameters, denoted $\lambda_{t,\mathcal{R}}$ to reflect their dependence on both O_t and $\mathcal{R}$, may be conveniently separated into face dependent parameters λ_t, i.e., parameters that are directly extracted from O_t, and face independent transformation parameters $\lambda_{\mathcal{R}}$, i.e., the parameters of the shared transformation model $\mathcal{R}$ which can be reliably trained by pooling together the training images of all available individuals.

A translation τ maps a position in O_q into another position in O_t, so that a feature vector $o_{i,j}$ in O_q needs to be matched to a feature vector in O_t that will be denoted $m^{\tau}_{i,j}$ (cf. Fig. 1). The emission probability $b^{\tau}_{i,j}$ represents the cost of matching $o_{i,j}$ and $m^{\tau}_{i,j}$. We model $b^{\tau}_{i,j}$ with a mixture of Gaussians. This choice is motivated by the fact that linear combinations of Gaussians can approximate arbitrarily shaped densities:

$$b^{\tau}_{i,j} = \sum_{k=1}^{K_{i,j}} w^k_{i,j} b^{\tau,k}_{i,j}. \tag{2}$$

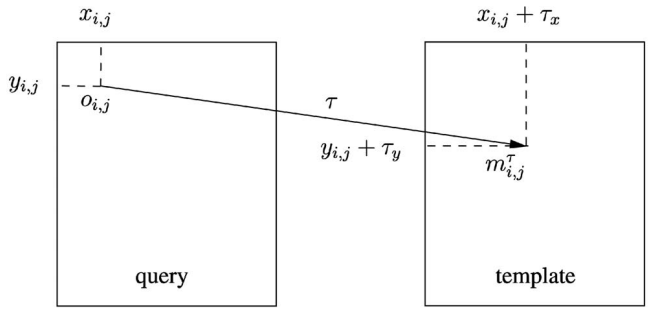


Fig. 1. Local mapping of a feature vector $o_{i,j}$ in the query image O_q into a feature vector $m^{\tau}_{i,j}$ in the template image O_t.

$K_{i,j}$ is the number of components at position (i,j), $b^{\tau,k}_{i,j}$s are the component densities and $w^k_{i,j}$s are the mixture weights and must satisfy the following constraint:

$$\sum_{k=1}^{K_{i,j}} w^k_{i,j} = 1\ ,\ \forall (i,j). \tag{3}$$

Each component density is a D-variate Gaussian function of the form:

$$b^{\tau,k}_{i,j}(o_{i,j}) = \frac{\exp\left\{-\frac{1}{2}(o_{i,j} - \mu^{\tau,k}_{i,j})^T \Sigma^{k(-1)}_{i,j}(o_{i,j} - \mu^{\tau,k}_{i,j})\right\}}{(2\pi)^{\frac{D}{2}}|\Sigma^k_{i,j}|^{\frac{1}{2}}}, \tag{4}$$

where $\mu^{\tau,k}_{i,j}$ and $\Sigma^k_{i,j}$ are, respectively, the mean and covariance matrix of the Gaussian, D is the dimensionality of the feature space and $|\cdot|$ denotes the determinant operator. This HMM is nonstationary as the Gaussian parameters depend on the position (i,j). This allows to weight automatically the different parts of the face during the scoring process.

We now relate $\mu^{\tau,k}_{i,j}$ to $m^{\tau}_{i,j}$ by writing $\mu^{\tau,k}_{i,j}$ as a function of $m^{\tau}_{i,j}$. If we consider the case of an affine transformation, then we write $\mu^{\tau,k}_{i,j}$ as:

$$\mu^{\tau,k}_{i,j} = W^k_{i,j}\zeta^{\tau}_{i,j}, \tag{5}$$

where $\zeta^{\tau}_{i,j} = \begin{bmatrix} 1 \\ m^{\tau}_{i,j} \end{bmatrix}$ is a vector of size $D+1$ and $W^k_{i,j}$ is a $D \times (D+1)$ matrix.

Interestingly, similar equations have been written in the field of automatic speech recognition (ASR) for the class of speaker adaptive training (SAT) algorithms [29]. Especially in [30], [31], the authors make use of "bipartite" models for the Gaussian means to separate variabilities. These models are made of two components: one models mostly the speaker dependent (SD) part of the acoustic variabilities and the other the residual speaker independent (SI) variabilities. The Gaussian means are written as a function f of the SD parameters where the parameters of f are SI, which is exactly what is expressed by (5).

It is interesting to understand the meaning of the previous equations and, especially, the impact of the separation of the HMM parameters into face dependent parameters and face independent transformation parameters. While the shape of $b^{\tau}_{i,j}$ depends only on the face independent transformation parameters $w^k_{i,j}$, $\Sigma^k_{i,j}$, and $W^k_{i,j}$,

its mean should be approximately centered around $m^{\tau}_{i,j}$, a face dependent parameter.

Intuitively, $b^{\tau}_{i,j}$ models the intraclass variability of the face around position (i,j).

3.2 Transition Probabilities

The neighborhood consistency of the transformation is ensured via the transition probabilities of the HMM. If we assume that the 2D HMM is a first order Markov process, the transition probabilities are of the form $P(q_{i,j}|q_{i,j-1}, q_{i-1,j})$. We would like to outline that the choice of this simple first order model is primarily motivated by its low complexity. However, an improved performance may be obtained with a richer model, by going beyond the first order statistics, at the expense of an increase of the computational cost.

We show in the next section that a 2D HMM can be approximated by a turbo hidden Markov model (T-HMM): a set of horizontal and vertical 1D HMMs that "communicate" through an iterative process. The transition probabilities of the corresponding horizontal and vertical 1D HMMs are denoted: $a^{\mathcal{H}}_{i,j}(\tau';\tau) = P(q_{i,j} = \tau'|q_{i,j-1} = \tau)$ and $a^{\mathcal{V}}_{i,j}(\tau';\tau) = P(q_{i,j} = \tau'|q_{i-1,j} = \tau)$.

Invariance to global shift in face images is a desirable property. Hence, if $\tau' = \tau + \delta\tau$, we choose $a^{\mathcal{H}}$ and $a^{\mathcal{V}}$ to be of the form:

$$a^{\mathcal{H}}_{i,j}(\tau + \delta\tau;\tau) = a^{\mathcal{H}}_{i,j}(\delta\tau), \tag{6}$$

$$a^{\mathcal{V}}_{i,j}(\tau + \delta\tau;\tau) = a^{\mathcal{V}}_{i,j}(\delta\tau). \tag{7}$$

We can apply further constraints on the transition probabilities to reduce the number of free parameters in our system. For instance, we can assume separable transition probabilities. If $\delta\tau = (\delta\tau_x, \delta\tau_y)$, then:

$$a^{\mathcal{H}}_{i,j}(\delta\tau) = a^{\mathcal{H}x}_{i,j}(\delta\tau_x) \times a^{\mathcal{H}y}_{i,j}(\delta\tau_y), \tag{8}$$

$$a^{\mathcal{V}}_{i,j}(\delta\tau) = a^{\mathcal{V}x}_{i,j}(\delta\tau_x) \times a^{\mathcal{V}y}_{i,j}(\delta\tau_y). \tag{9}$$

We can also assume parametric transition probabilities. If O_t and O_q have the same scale and orientation, then the horizontal transition probabilities could have the following form:

$$a^{\mathcal{H}}_{i,j}(\delta\tau) \propto \exp\left\{-\frac{1}{2}\left[\left(\frac{\delta\tau_x}{\sigma^{\mathcal{H}x}_{i,j}}\right)^2 + \left(\frac{\delta\tau_y}{\sigma^{\mathcal{H}y}_{i,j}}\right)^2\right]\right\}. \tag{10}$$

A similar formula can be derived for vertical transition probabilities. Another idea to reduce the number of transition probability parameters would be to use the face symmetry.

$a^{\mathcal{H}}_{i,j}$ and $a^{\mathcal{V}}_{i,j}$ model, respectively, the horizontal and vertical elastic properties of the face at position (i,j) and are part of the face transformation model $\mathcal{R}$. Note that using multiple horizontal and vertical transition probabilities at different locations enables to model the different elastic properties of the various parts of the face.

Finally, we consider the remaining set of HMM parameters—the initial occupancy probabilities. We assume herein that the initial occupancy probability distribution is uniform, to ensure invariance to global translations of face images.

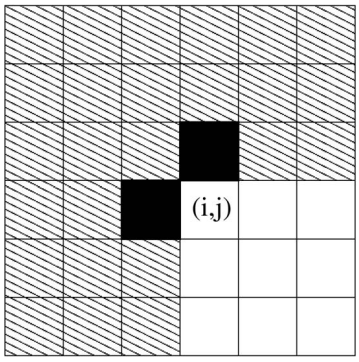


Fig. 2. Markovian property of transitions among states for the first order Markov mesh random field (cf. (11)).

3.3 Turbo Hidden Markov Models

The one-dimensional hidden Markov model (1D HMM) is a class of stochastic signal model which has a long history of success in various problem domains, perhaps most notably in speech recognition. This success is largely due to the development of computationally efficient algorithms to solve the three fundamental problems of HMM design, namely, the forward-backward, Viterbi and Baum-Welch algorithms [32].

The Markov random field (MRF) is the 2D counterpart of the 1D Markov chain where the natural ordering of past, present and future is replaced by the spatial concept of neighborhood. The MRF modeling process generally consists of the following steps [33]: defining a neighborhood system, defining cliques, defining the prior clique potentials, deriving the likelihood energy, and deriving the posterior energy. In this paper, we consider a subclass of MRF models, the Markov mesh random field (MMRF), which reintroduces the notion of past, present, and future thanks to the raster scan [34]. In this paper, we focus on the first order MMRF. Let $Q = \{q_{i,j}, i = 1,\ldots,I, j = 1,\ldots,J\}$ be a $I \times J$ array of states and let $Q_{i,j}$ be the set of states to the left or above $q_{i,j}$: $Q_{i,j} = \{q_{m,n}, m < i \text{ or } n < j\}$. Then the first order MMRF can be defined by the following property (cf. also Fig. 2):

$$P(q_{i,j}|Q_{i,j}) = P(q_{i,j}|q_{i,j-1}, q_{i-1,j}). \tag{11}$$

Reintroducing the notion of past and future is beneficial as it allows to develop the joint distribution of states $P(Q|\lambda)$ as is the case for the 1D HMM. Thus, the forward-backward, Viterbi and Baum-Welch algorithms developed for the 1D HMM can be extended to the 2D case. However, even with the simple first-order Markovian model considered, the direct extension of these algorithms to the 2D case is exponential in the size of the data [35], and hence intractable for most applications of practical value. Thus, approximations are required.

Many approximations of the 2D HMM were suggested. It seems that most approaches attempt to replace the 2D HMM with a 1D HMM [34], [36] or a set of 1D HMMs [35], [37], [38], [39] whose properties are well understood. In [39], the turbo hidden Markov model (T-HMM) was introduced, in

reference to the celebrated turbo error-correcting codes, as an efficient approximation of the 2D HMM. A T-HMM consists of a set of horizontal and vertical 1D HMMs that "communicate" through an iterative process by inducing prior probabilities on each other. Thus, the use of the T-HMM corresponds to a simple elasticity model of the face where each part of the face is linked to its horizontal and vertical neighbors by springs. The T-HMM framework provides efficient approximate formulas to 1) compute the score $P(O_q|O_t,\mathcal{R})$ and 2) estimate the parameters of $\mathcal{R}$.

3.3.1 Estimation of $P(O_q|O_t,\mathcal{R})$

The computation of $P(O_q|O_t,\mathcal{R})$ is based on a modified version of the forward-backward algorithm which is applied successively and iteratively on the rows and columns until the horizontal and vertical priors reach a desired level of agreement. This algorithm is linear in the size of the data modulo the number of iterations. For more details the reader can refer to [39].

3.3.2 Parameter Estimation

We recall that the parameters to be estimated are the $\lambda_{\mathcal{R}}$ parameters, i.e., the parameters of the face transformation model: $w_{i,j}^k$, $W_{i,j}^k$, $\Sigma_{i,j}^k$, $a_{i,j}^{\mathcal{H}}$, and $a_{i,j}^{\mathcal{V}}$. During training, we present pairs of images (O_t^p, O_q^p) that belong to the same person and optimize the transformation parameters $\lambda_{\mathcal{R}}$, to increase the likelihood value $\prod_p P(O_q^p|O_t^p,\lambda_{\mathcal{R}})$ (Maximum Likelihood (ML) estimation).

We recall that Q denotes a sequence of states: $Q=\{q_{i,j}, i=1,\ldots,I, j=1,\ldots,J\}$ and $\lambda_{\mathcal{R}}$ the current estimate of the HMM parameters. The re-estimation formulas for HMM parameters are usually derived by maximizing Baum's auxiliary function:

$$\mathcal{Q}(\lambda'_{\mathcal{R}}|\lambda_{\mathcal{R}}) = \sum_Q P(Q|O_q,O_t,\lambda_{\mathcal{R}}) \log P(O_q,Q|O_t,\lambda'_{\mathcal{R}}). \quad (12)$$

with respect to $\lambda'_{\mathcal{R}}$. It has been proven that the maximization of $\mathcal{Q}(\lambda'_{\mathcal{R}}|\lambda_{\mathcal{R}})$ leads to an increased likelihood (see, e.g., [40]). As noted in [32], the re-estimation formulas can be interpreted as an implementation of the Expectation-Maximization (EM) algorithm in which the E step is the calculation of the auxiliary function $\mathcal{Q}(\lambda'_{\mathcal{R}}|\lambda_{\mathcal{R}})$, and the M step is the maximization over $\lambda'_{\mathcal{R}}$.

Let us consider the case where we have one pair of images (O_t, O_q). During the E-step, one performs a modified forward-backward to estimate $\gamma_{i,j}^{\tau} = P(q_{i,j}=\tau|O_q,O_t,\lambda_{\mathcal{R}})$ (occupancy probability),

$$\xi_{i,j}^{\mathcal{H}}(\tau+\delta\tau,\tau) = P(q_{i,j+1}=\tau+\delta\tau, q_{i,j}=\tau|O_q,O_t,\lambda_{\mathcal{R}})$$

and

$$\xi_{i,j}^{\mathcal{V}}(\tau+\delta\tau,\tau) = P(q_{i+1,j}=\tau+\delta\tau, q_{i,j}=\tau|O_q,O_t,\lambda_{\mathcal{R}}).$$

We also define $\gamma_{i,j}^{\tau,k}$, the probability of being in state $q_{i,j}=\tau$ at position (i,j) with the kth mixture component accounting for $o_{i,j}$, which can be estimated as follows:

$$\gamma_{i,j}^{\tau,k} = \gamma_{i,j}^{\tau} \frac{w_{i,j}^k b_{i,j}^{\tau,k}}{\sum_{k=1}^{K_{i,j}} w_{i,j}^k b_{i,j}^{\tau,k}}. \quad (13)$$

During the M-step, we set $\frac{\delta\mathcal{Q}}{\delta w_{i,j}^k}=0$, $\frac{\delta\mathcal{Q}}{\delta W_{i,j}^k}=0$, $\frac{\delta\mathcal{Q}}{\delta \Sigma_{i,j}^k}=0$, $\frac{\delta\mathcal{Q}}{\delta a_{i,j}^{\mathcal{H}}}=0$, and $\frac{\delta\mathcal{Q}}{\delta a_{i,j}^{\mathcal{V}}}=0$ and obtain, respectively, the following reestimation formulas:

$$\hat{w}_{i,j}^k = \frac{\sum_\tau \gamma_{i,j}^{\tau,k}}{\sum_\tau \gamma_{i,j}^{\tau}}, \quad (14)$$

$$\hat{W}_{i,j}^k = \sum_\tau \left(\gamma_{i,j}^{\tau,k} o_{i,j} \zeta_{i,j}^{\tau\,T}\right) \left(\sum_\tau \gamma_{i,j}^{\tau,k} \zeta_{i,j}^{\tau} \zeta_{i,j}^{\tau\,T}\right)^{-1}, \quad (15)$$

$$\hat{\Sigma}_{i,j}^k = \frac{\sum_\tau \gamma_{i,j}^{\tau,k} \left(o_{i,j} - \hat{W}_{i,j}^k \zeta_{i,j}^{\tau}\right)\left(o_{i,j} - \hat{W}_{i,j}^k \zeta_{i,j}^{\tau}\right)^T}{\sum_\tau \gamma_{i,j}^{\tau,k}}, \quad (16)$$

$$\hat{a}_{i,j}^{\mathcal{H}}(\delta\tau) = \frac{\sum_\tau \xi_{i,j}^{\mathcal{H}}(\tau+\delta\tau,\tau)}{\sum_\tau \gamma_{i,j}^{\tau}}, \quad (17)$$

$$\hat{a}_{i,j}^{\mathcal{V}}(\delta\tau) = \frac{\sum_\tau \xi_{i,j}^{\mathcal{V}}(\tau+\delta\tau,\tau)}{\sum_\tau \gamma_{i,j}^{\tau}}. \quad (18)$$

Estimating $W_{i,j}^k$ requires to solve a linear system of $D+1$ equations with $D+1$ unknowns (cf. (15)). In (16), we assumed the general case of full covariance matrices. In (17) and (18), we assumed unconstrained transition probabilities. While these equations are given for the improbable case where HMM parameters are estimated with only one pair of images, their extension to the case of multiple pairs of images is straightforward.

4 Modeling Illumination Variations

While grid transformations are useful to compensate for facial expressions and, as we will show experimentally, for pose variations, they are of no use when dealing with illumination variations. Hence, new transformations (i.e., states) should be allowed by our HMM.

In Section 4.1, we first show how to transform the illumination into an additive component in the feature domain and how to modify the Gaussian parameters of the emission probabilities to account for it. In Section 4.2, we explain how to constrain variation in illumination through the transition probabilities. Finally, in Section 4.3, we briefly introduce the turbo state-space model (T-SSM) which is the counterpart of the T-HMM in the case where the states are continuous variables. The T-SSM framework provides efficient formulas to 1) estimate the best sequence of "illumination" states and 2) estimate the parameters of $\mathcal{R}$ which correspond to the illumination compensation part of our algorithm.

4.1 Modeling Illumination

The starting point for illumination modeling is the well-known assumption that an image I can be seen as the product of a reflectance R and an illumination L [41]:

$$I(x,y) = R(x,y) \times L(x,y). \quad (19)$$

Applying the logarithm operator, we obtain:

$$\log I(x,y) = \log R(x,y) + \log L(x,y). \quad (20)$$

and the illumination turns into an additive term in the pixel domain. If the feature extraction operator $\mathcal{F}$ is linear, such as convolution, then we obtain:

$$\mathcal{F}\{\log I(x,y)\} = \mathcal{F}\{\log R(x,y)\} + \mathcal{F}\{\log L(x,y)\}. \quad (21)$$

and illumination remains additive in the feature domain.

The idea is hence to introduce feature transformations to model the illumination and to enforce consistency between feature transformations at adjacent positions, in the same manner we enforced consistency between grid transformations, to constrain the illumination variation. Hence, our states which represent both local grid and feature transformations are now doubly indexed: $q_{i,j} = (\tau_{i,j}, \phi_{i,j})$. $\tau_{i,j}$ and $\phi_{i,j}$ are, respectively, the grid and feature transformation parts of the state. If $q_{i,j} = (\tau, \phi)$, the emission probability $b_{i,j}^{\tau,\phi}$ is still modeled with a mixture of Gaussians:

$$b_{i,j}^{\tau,\phi} = \sum_{k=1}^{K_{i,j}} w_{i,j}^k b_{i,j}^{\tau,\phi,k}, \quad (22)$$

where the $b_{i,j}^{\tau,\phi,k}$s are D-variate Gaussians with means $\mu_{i,j}^{\tau,\phi,k}$ and covariance matrices $\Sigma_{i,j}^k$. If the "feature" state ϕ also denotes the additive contribution of the illumination in the feature domain, the Gaussian means are of the form:

$$\mu_{i,j}^{\tau,\phi,k} = \mu_{i,j}^{\tau,k} + \phi = W_{i,j}^k \zeta_{i,j}^{\tau} + \phi. \quad (23)$$

4.2 Constraining the Illumination Variation

If we assume that grid and feature transformations model, respectively, differences in facial expression and illumination between images, and that facial expression and illumination *variations* are mostly independent (i.e., a facial expression change between two adjacent positions has a limited impact on the illumination change between the same positions and vice versa), then the horizontal and vertical transition probabilities can be separated as follows:

$$P(q_{i,j}|q_{i,j-1}) = P(\tau_{i,j}|\tau_{i,j-1}) \times P(\phi_{i,j}|\phi_{i,j-1}), \quad (24)$$

$$P(q_{i,j}|q_{i-1,j}) = P(\tau_{i,j}|\tau_{i-1,j}) \times P(\phi_{i,j}|\phi_{i-1,j}). \quad (25)$$

While the choice of a discrete number of grid transformations is natural due to the discrete nature of the feature extraction grid of the template image, it is easier to deal with the illumination with an *infinite continuous* set of illumination states. We choose the horizontal and vertical illumination components of the transition probabilities to be D-variate Gaussians:

$$P(\phi_{i,j} = \phi|\phi_{i,j-1} = \phi') = P(\phi_{i,j} = \phi|\phi_{i-1,j} = \phi') = \frac{\exp\left\{-\frac{1}{2}(\phi - \phi')^T S^{(-1)}(\phi - \phi')\right\}}{(2\pi)^{\frac{D}{2}}|S|^{\frac{1}{2}}}. \quad (26)$$

The choice of such a transition probability is primarily motivated by its computational tractability. To reduce even more the complexity, in the following we assume that the covariance matrix S is diagonal and, therefore, that the components of the feature vectors are independent from each other. S is the only parameter of our illumination transformation model and it models the speed of variation of the illumination in each of the feature components.

4.3 Turbo State-Space Models

An HMM with an infinite continuous set of states is generally referred to as a state-space model (SSM) [42]. The growth in complexity that plagues the 2D HMM also arises in the case of the 2D SSM and approximations are required. In [43], the T-HMM framework was extended to the continuous state turbo SSM (T-SSM). The T-SSM provides efficient approximate formulas to the two following problems: 1) estimate the best sequence of "illumination" states and 2) estimate the parameters of $\mathcal{R}$ which correspond to the illumination compensation part of our algorithm, i.e., the diagonal matrix S.

We denote $Q = (T, \Phi)$ a sequence of states where T is a sequence of grid states: $T = \{\tau_{i,j}, 1 \le i \le I, 1 \le j \le J\}$ and Φ is a sequence of feature states: $\Phi = \{\phi_{i,j}, 1 \le i \le I, 1 \le j \le J\}$. In the case where we attempt to model facial expressions and illumination variations, our similarity measure between face images is:

$$P(O_q|O_t, \Phi^*, \mathcal{R}), \quad (27)$$

where Φ^* is the sequence of feature states that best explains the illumination variation.

4.3.1 Finding the Best Sequence of States

In the case where the system can be in only one grid state at each position and where emission probabilities are Gaussian, we can extend the modified forward-backward introduced for the T-HMM to the T-SSM (see [43]). During the modified forward-backward, we can estimate $\gamma_{i,j}^{\phi} = P(q_{i,j} = \phi|O_q, O_t, \lambda_{\mathcal{R}})$ and then choose the sequence of locally optimal states:

$$\phi^* = \arg\max_{\phi} \gamma_{i,j}^{\phi}. \quad (28)$$

Although choosing the sequence of locally optimal states may not lead to the sequence of globally optimal states, this approximation is valid in the case where the best sequence of states accounts for most of the total probability.

In the case where we perform an elastic matching and where emission probabilities are mixtures of Gaussians, a direct application of the modified forward-backward would be too computationally intensive. Instead we propose to apply iterative passes to find successively the grid states, Gaussian indexes and feature states that best explain the transformations between two images. Let $k_{i,j}$ be the Gaussian index in the emission probability at position (i, j) and let K be a sequence of Gaussian indexes: $K = \{k_{i,j}, 1 \le i \le I, 1 \le j \le J\}$. Let us also denote, respectively, T_n, K_n, and Φ_n the best set of grid states, Gaussian indexes and illumination states after the nth iteration.

The iterative procedure is as follows:

1. Initialize Φ_0: $\forall(i,j)$, $\phi_{i,j} = 0$, i.e., we assume that there is no illumination variation between O_q and O_t.
2. $T_n = \arg\max_T \log P(T|O_q, O_t, \Phi_{n-1}, \lambda|\mathcal{R})$: During the forward-backward, one estimates the occupancy

probabilities $\gamma_{i,j}^{\tau}$ and chooses at each position (i,j) the state $\tau_{i,j} = \tau^*$ such that: $\tau^* = \arg\max_\tau \gamma_{i,j}^{\tau}$.

3. $K_n = \arg\max_K \log P(K|O_q, O_t, T, \Phi_{n-1}, \lambda_{\mathcal{R}})$: During the forward-backward, one can also estimate $\gamma_{i,j}^{\tau,k}$. If τ^* is the optimal state found at position (i,j), then the optimal Gaussian index $k_{i,j} = k^*$ is chosen such that $k^* = \arg\max_k \gamma_{i,j}^{\tau^*,k}$.
4. $\Phi_n = \arg\max_\Phi \log P(\Phi|O_q, O_t, T_n, K_n, \lambda_{\mathcal{R}})$: We apply the modified forward-backward for the T-SSM and estimate $\gamma_{i,j}^{\phi}$ and choose at each position (i,j) the state $\phi_{i,j} = \phi^*$ such that: $\phi^* = \arg\max_\phi \gamma_{i,j}^{\phi}$.
5. Go back to Step 2 until T_n, K_n, and Φ_n converge.

Although this iterative procedure is not guaranteed to converge, it does provide acceptable results as shown in the experimental results section. Note that, once Φ^* is obtained, the computation of the score $P(O_q|O_t, \Phi^*, \lambda_{\mathcal{R}})$ is done with the modified forward-backward for the T-HMM except that we replace the features $o_{i,j}$ with their illumination compensated version $o_{i,j} - \phi_{i,j}^*$.

4.3.2 Parameter Estimation

The estimation of the parameter $S = \mathrm{diag}\{s[1] \dots s[D]\}$ is performed through the maximization of Baum's auxiliary function as done in Section 3.3.2. This maximization can be done independently per feature components and, for the considered transition probabilities, this leads to the following re-estimation formula:

$$\hat{s}^2 = \frac{\sum_{i,j} \int_{\phi,\phi'} (\phi - \phi')^2 [\xi_{i,j}^{\mathcal{H}}(\phi,\phi') + \xi_{i,j}^{\mathcal{V}}(\phi,\phi')] d\phi d\phi'}{(I-1) \times J + I \times (J-1)}. \quad (29)$$

Note that there exists a closed form formula, which is not shown here due to its complexity.

It is interesting to note a very recent approach within the discipline of speech recognition [44], [45], which employs a philosophy that bears some similarity to our work on illumination compensation, although in an altogether different field and context. Indeed, in [44], the noise is modeled as a sequence of states of a dynamical system with a continuum of states. Observations generated by such a system are assumed to be related to the state of the system by a functional relation which models clean speech as the corrupting influence of noise. In our case, we assume that variations due to facial expressions corrupt the illumination signal. The work of [45] brings important differences, one of which is to perform a joint noise and speech tracking. This is fairly similar with the joint grid and feature transformations estimation used in our approach.

5 Related Work

As explained in the introductory section of this paper, only a few face recognition algorithms concentrate on computing a distance between face images. A comprehensive review of the literature on face recognition is beyond the scope of this paper and the interested reader is referred to [3], [4], [46]. In this section, we will focus on the Bayesian intra/extrapersonal criterion [17], which will be referred to as BAID as the basic idea of this approach is to perform a Bayesian Analysis of Image Differences. There are two main reasons for this choice. It was one of the top performers during the 1996 FERET evaluations [5] (and remains one of the most successful face recognition algorithms to date [47]) and it can be related to our approach. Hence, in Section 5.1, we first briefly describe the BAID algorithm and then show in Section 5.2 that the BAID and the proposed probabilistic model of face mapping can be understood as different (competing) approximations of the same high-dimensional density.

5.1 The BAID Algorithm

The focus of [17] is on modeling the difference Δ between face images. The observed variability can be explained by two mutually exclusive classes of variability: the intrapersonal variability Ω_I (equivalent to our notation $\mathcal{R}$) and the extrapersonal variability Ω_E. The chosen measure of similarity between two face images is $P(\Omega_I|\Delta)$ which, using Bayes rule, can be evaluated as follows:

$$P(\Omega_I|\Delta) = \frac{P(\Delta|\Omega_I)P(\Omega_I)}{P(\Delta|\Omega_I)P(\Omega_I) + P(\Delta|\Omega_E)P(\Omega_E)}. \quad (30)$$

A simple ML formulation, which uses only the intrapersonal variability, is often preferred to the previous MAP classifier, as it reduces the computation by a factor of two at the cost of very little degradation of the performance. In such a case, the similarity score is simply $P(\Delta|\Omega_I)$.

The difference between face images of the same person is assumed to be a normally distributed random variable:

$$P(\Delta|\Omega_I) = \frac{\exp\{-\frac{1}{2}\Delta^T S^{-1} \Delta\}}{(2\pi)^{N/2}|S|^{1/2}}. \quad (31)$$

Due to the high dimensionality of Δ (e.g., for 128×128 pixels images, $N = 16,384$) the direct estimation of the parameter of this probability density function, i.e., of the covariance matrix S, is difficult. Moreover, estimating $P(\Delta|\Omega_I)$ can be very computationally intensive. Therefore, the intrapersonal image difference space is separated into a principal subspace F and its orthogonal complement $\bar{F}$. Thus, $P(\Delta|\Omega_I)$ can be approximated as the product of two terms:

$$P(\Delta|\Omega_I) \approx \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{M} \frac{y_i^2}{\lambda_i}\right\}}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}} \frac{\exp\left\{-\frac{1}{2}\frac{\epsilon^2(\Delta)}{\rho}\right\}}{(2\pi\rho)^{(N-M)/2}}, \quad (32)$$

where y_i is the projection of Δ on the ith principal direction of F, λ_i is the eigenvalue associated with this direction, $\epsilon^2(\Delta)$ is the squared Euclidean distance of Δ to F, and ρ is the average eigenvalue in $\bar{F}$ (see [16], [17], [18] for more details).

5.2 Relationship between BAID and the Proposed Approach

We can now relate this approach to the proposed framework. First, we could envision applying the Gaussian classifier (and, thus, BAID) not on the pixel to pixel difference between face images but on the difference between their representations, i.e., after a feature extraction step. We assume that the set of feature vectors $\{o_{i,j}\}$ and $\{m_{i,j}\}$ are extracted on a grid, respectively, from the query image O_q and the template image O_t. Note that [16] considers the case where feature vectors are nonoverlapping gray-level blocks that cover the whole

image. We denote $\Delta_{i,j} = o_{i,j} - m_{i,j}$. If we assume that the difference between feature vectors at adjacent positions is uncorrelated, then the covariance matrix S is block diagonal. If $S_{i,j}$ is the block corresponding to position (i,j) and if D is the dimension of the feature vectors, then:

$$P(\Delta|\Omega_I) = \prod_{i,j} P(\delta_{i,j}|\Omega_I) \tag{33}$$

$$= \prod_{i,j} \frac{\exp\left\{-\frac{1}{2}\delta_{i,j}^T S_{i,j}^{-1}\delta_{i,j}\right\}}{(2\pi)^{D/2}|S_{i,j}|^{1/2}}. \tag{34}$$

This corresponds to the probabilistic distance for our classifier in the very simple case where we perform a rigid matching (no grid transformation) without any illumination compensation (no feature transformation) and where we use a single Gaussian per mixture with a full covariance matrix. Hence, while BAID and a simplified version of our algorithm can be viewed as different approximations of the same high dimensional density, these two algorithms pursue radically different approaches. BAID uses a *global* approach as the difference between faces is modeled in its entirety. The proposed approach makes use of a *feature-based* approach as we consider local representations of the face.

6 Experimental Results

In this section, we present an experimental comparison of BAID and our probabilistic mapping with local transforms (PMLT). In Section 6.1, we first briefly introduce the four databases used to carry out our experiments. In Section 6.2, we describe the features extracted from the face images and which are used by the face classifiers. The training procedures employed in the design of the BAID and PMLT classifiers are specified in Section 6.3. Finally, we carry out the comparison and evaluate the robustness of BAID and PMLT in the case of a degradation of the image resolution, an imprecise segmentation and a variation in facial expression, illumination or pose Section 6.4. All the results we present are for *identification* experiments.

6.1 The Databases

The Facial Recognition Technology (FERET) database [5] contains over 14,000 images taken from 1,199 individuals. For each individual, two frontal views were taken (FA and FB images) and a different facial expression was requested for the second frontal image. For 200 individuals, a third frontal image was taken with a different camera and different lighting (FC images) and a set of images was collected at various aspects ranging from right lo left profile. For some individuals, a second set of images was taken on a later date (duplicate sets).

6.1.1 Yale B

The Yale face database B (Yale B) [23] contains 5,850 images of 10 subjects. Some variation across pose was obtained by taking pictures simultaneously with nine cameras. To get wide illumination variations, the database was captured using a purpose-built illumination rig with 64 strobes. The 64 images of a subject in a particular pose were acquired in about 2 seconds, so there is only small change in head pose and facial expression for those 64 images. An additional set of images was captured with no strobe going off (ambient lighting).

6.1.2 PIE

The CMU Pose Illumination Expression (PIE) database [22] contains over 40,000 images taken from 68 individuals. To obtain large variations across pose, a set of 13 cameras was used. To obtain significant illumination variations, a flash system similar to the one constructed at the Yale university was used. The flash system consisted of 21 flashes. Since images were captured with and without background lighting and since one picture was taken with ambient lighting, $21 \times 2 + 1 = 43$ different illumination conditions were obtained.

6.1.3 AR

The Alex Martínez-Robert Benavente (AR) face database [24] contains over 4,000 images of 126 subjects. Images feature frontal view faces with different facial expressions (neutral, smile, anger, scream), illumination conditions (left light on, right light on, both lights on) and occlusions (wearing sun glasses, wearing a scarf). Each person participated in two sessions separated by two weeks and the same set of pictures were taken in both sessions.

For all images, we manually located the position of the eyes and the nose and we extracted normalized 128×128 pixels facial images. It was shown in [18] that BAID performed very well even for face images with a much coarser resolution (down to 21×12 pixels). Therefore, we will carry out a set of experiments in Section 6.4.1 to know how the proposed PMLT depends on the image resolution. Also, in Section 6.4.2, we will evaluate the impact of an inaccurate location of facial features and, thus, of an imprecise segmentation.

6.2 Features

As a preprocessing step, we first applied a log in the pixel domain to partially compensate for illumination effects [48], [49]. We then extracted Gabor features which have long been successfully applied to face recognition [20] and facial analysis [50]. Gabor wavelets are plane waves restricted by a Gaussian envelope.

To define a bank of Gabor wavelets, [51] suggests to partition the spectral half plane into M frequency and N orientation bands. The set of filters is defined as follows in the Fourier domain:

$$G_{i,j}(\omega_u, \omega_v) = \exp\left\{-\frac{1}{2}\left[\frac{\omega_u^2}{\sigma_{\rho_i}^2} + \frac{\omega_v^2}{\sigma_{\theta_i}^2}\right]\right\} \tag{35}$$
$$i = 1, \ldots, M, j = 1, \ldots, N$$

with:

$$\begin{pmatrix}\omega_u \\ \omega_v\end{pmatrix} = \begin{bmatrix}\cos(\omega_{\theta_j}) & \sin(\omega_{\theta_j}) \\ -\sin(\omega_{\theta_j}) & \cos(\omega_{\theta_j})\end{bmatrix}\begin{pmatrix}\omega_x \\ \omega_y\end{pmatrix} - \begin{pmatrix}\omega_{\rho_i} \\ 0\end{pmatrix}. \tag{36}$$

ω_{ρ_i} and σ_{ρ_i} are, respectively, the radial center and bandwidth and ω_{θ_j} and σ_{θ_i} are, respectively, the angular center and bandwidth. These parameters are defined as follows:

$$\omega_{\rho_i} = \omega_{min} + \sigma_0 \frac{(f+1)f^{i-1} - 2}{f-1}, \quad (37)$$

$$\sigma_{\rho_i} = \sigma_0 f^{i-1}, \quad (38)$$

$$\omega_{\theta_j} = \frac{(j-1)\pi}{N}, \quad (39)$$

$$\sigma_{\theta_i} = \frac{\pi\omega_{\rho_i}}{2N}, \quad (40)$$

with σ_0 given by:

$$\sigma_0 = \frac{\omega_{max} - \omega_{min}}{2}\left(\frac{f-1}{f^M - 1}\right). \quad (41)$$

Therefore, to define a bank of Gabor wavelets, one has to set five parameters: ω_{min}, ω_{max}, f, M, and N. After preliminary experiments, we chose $\omega_{min} = \pi/24$, $\omega_{max} = \pi/3$, $f = \sqrt{2}$, $M = 4$, and $N = 6$, which resulted in 24 dimensional feature vectors. Gabor responses are obtained through the convolution of an image and the Gabor wavelets. We use the modulus of these responses as feature vectors which introduces a nonlinearity in the computation of our features. Thus, the illumination cannot be considered as a perfectly additive term in the feature domain.

As the focus is on the comparison of the BAID and PMLT classifiers, we used the same features for both approaches. For the PMLT, feature vectors were extracted every 16 pixels of the query images and every four pixels of the template images, thus limiting the precision of a local grid transformation to four pixels in both horizontal and vertical directions. For the BAID algorithm, feature vectors were extracted every four pixels for both the template and query images.

6.3 Training

The training was performed on a set of 500 persons extracted from the FERET database. These 500 persons have one FA and one FB image and 200 of them have an additional FC image. Hence, training was performed with $500 \times 2 + 200 = 1,200$ images. We now describe the training of the PMLT and BAID classifiers.

6.3.1 PMLT

The design of the PMLT requires a number of choices. First, we chose the covariance matrices $\Sigma_{i,j}$ to be diagonal as a linear combination of Gaussians with diagonal covariances can approximate any distribution with arbitrary precision and as diagonal covariance matrices require significantly less computation than full covariance matrices. As for the matrix $W^k_{i,j}$, it can be separated into $W^k_{i,j} = (\delta^k_{i,j} \vdots \Pi^k_{i,j})$ where $\delta^k_{i,j}$ is a vector of size D and $\Pi^k_{i,j}$ is a $D \times D$ matrix. Thus, the product $W^k_{i,j}\zeta^\tau_{i,j}$ takes the form $\Pi^k_{i,j}m^\tau_{i,j} + \delta^k_{i,j}$. We tried the following possibilities:

- $\Pi^k_{i,j} = I_D$, where I_D is the identity matrix of size D, i.e., we model only additive variabilities.
- $\Pi^k_{i,j}$ is diagonal and $\delta^k_{i,j} = [0 \cdots 0]^T$, i.e., we model only multiplicative variabilities.
- $\Pi^k_{i,j}$ is diagonal, i.e., we model both additive and multiplicative variabilities.

None of the systems seemed to clearly outperform the other ones for all conditions and we chose to model only additive variabilities as is the case for the BAID. Finally, we used general transition probabilities and to reduce the number of parameters to estimate, we used the face symmetry.

The model of face transformation is trained in two steps.

- During the first step, we train Gaussian parameters and transition probabilities $a^{\mathcal{H}}_{i,j}$ and $a^{\mathcal{V}}_{i,j}$. We assume that there is no illumination variation and fix at each position $(i,j)\phi_{i,j} = [0 \cdots 0]^T$. Therefore, this first part of the training only involves the FA and FB images. At each location (i,j), we start with one Gaussian per mixture (Gpm). We initially set $\delta_{i,j} = [0 \cdots 0]^T$ which is intuitive as, with this choice, $b^\tau_{i,j}(o_{i,j})$ is maximum if $o_{i,j} = m^\tau_{i,j}$. We first perform a rigid matching between the template and query images and estimate the covariance matrices $\Sigma_{i,j}$. As for the transition probabilities, they are initialized uniformly. Then, covariance matrices and transition probabilities are reestimated using iterative passes of the Baum-Welch algorithm until the likelihood of the training set converges. To train multiple Gpm, we used an iterative strategy inspired by the vector quantization (VQ) algorithm [52]. If a Gaussian was estimated with a sufficient number of observations, then it is split by introducing a small perturbation in $\delta^k_{i,j}$. Then, parameters are reestimated using the Baum-Welch algorithm. The splitting/retraining operations are repeated until the desired number of Gaussians is obtained. The maximum number of Gpm was set to 16 throughout our experiments.
- During the second step, we train only the S parameter to compensate for illumination variations and, thus, we also use the FC images. To train S, we started with the previously well-trained system. We initialized S in the following manner $S = sI_D$ with s very large and reestimated the matrix with the Baum-Welch algorithm.

6.3.2 BAID

The ML classifier was trained exactly as described in [16] with the 1,200 available images and, after preliminary experiments, we decided to keep $E = 100$ eigenvectors. We tried to improve the performance by modeling extrapersonal differences, i.e., by using the MAP classifier. However, for our set of experiments, we did not observe any significant difference between the ML and MAP classifiers. These results are consistent with findings from other researchers as it was even shown in [53] that the ML classifier could lead to a slightly better performance than the more complex MAP classifier. The reason for the very small observed difference between the ML and MAP classifiers is explained in [54]. As the extrapersonal subspace is similar to the PCA eigenspace, it does not contribute much to separating intra and extrapersonal variabilities.

TABLE 1
Influence of the Image Resolution

resolution (in pixels)	BAID	PMLT
128×128	92%	93%
64×64	92%	93%
32×32	91%	92%
16×16	87%	88%

Results on the FERET database.

TABLE 2
Imprecise Segmentation Results on the FERET Database

σ (in pixels)	BAID	PMLT
0	92%	93%
1	91%	93%
2	87%	**92%**
3	83%	**90%**

6.4 Results

We performed six sets of experiments. We first evaluated the robustness of BAID and PMLT with respect to a degradation of the image resolution or an imprecise segmentation of the face. We then assessed their performance in the presence of facial expression, illumination, or pose variations. In the last set of experiments, we evaluated their performance in the challenging case of both illumination and pose variations. For each set of experiments, we carried out the tests on the database(s) that, we thought, would be the most interesting for the considered variability.

For all the following experiments, we performed McNemar's test to determine whether the observed difference in performance between the BAID and the PMLT could be considered significant. If one of the two classifiers outperforms with more than 95 percent confidence the other classifier, then its score is bold-faced.

6.4.1 Image Resolution

The robustness of BAID and PMLT with respect to the image resolution was evaluated on the FERET database. The test data consisted of 695 persons who were not already in the training set. FA images were used as enrollment/gallery data and FB images as test/probe data. Results are presented in Table 1.

We can see that the dependence of BAID and PMLGT on the image resolution is similar. Indeed, there is little degradation of the performance down to 32×32 pixels and a significant degradation for 16×16 pixels.

6.4.2 Imprecise Segmentation

The robustness of BAID and PMLT with respect to an imprecise segmentation was evaluated on the same FERET data set as in the previous experiment. To simulate an imprecise segmentation at test time, the localization of facial features on query images was perturbated with an additive Gaussian noise with mean zero and a varying standard deviation σ. The localization of facial features for enrollment images was not perturbated. The rational behind this choice is the fact that enrollment is often supervised and, thus, an incorrect segmentation can be manually corrected.

Results are presented in Table 2. Obviously, PMLT is much more robust to an imprecise segmentation than BAID. We believe that the robustness of PMLT is due to the local grid transformations which allow more flexibility in the matching. Therefore, we forced the PMLT to perform a rigid matching by constraining the system to be at each position (i, j) in the state $\tau_{i,j} = (0, 0)$. The results we obtained with this rigid PMLT for a standard deviation of 1, 2, and 3 pixels were 90, 86, and 82 percent, respectively, which validates our claim.

6.4.3 Facial Expressions

The robustness of BAID and PMLT with respect to facial expressions was evaluated on the AR database. All the available persons were used. The image labeled 01, which corresponds to the neutral expression, was used as enrollment data and the images 02, 03, and 04, which correspond respectively to the smile, anger, and scream expressions, are used as test images (see Fig. 3).

Results are presented in Table 3. PMLT outperforms BAID for all expressions. Both BAID and PMLT perform fairly poorly for extreme facial expressions such as the scream of the AR database. Note however that this is not surprising as the training data, which contains only images from the FERET database, does not exhibit such variability.

We believe that the main reason for the difference in performance between the BAID and PMLT is the fact that grid transformations allow the PMLT to perform an elastic matching of facial images while BAID works directly on image differences and, thus, performs a rigid matching. To test this hypothesis, we forced the PMLT to perform a rigid matching as was done in the previous experiment. The results for this rigid PMLT are respectively 94, 89, and 66 percent for the smile, anger, and scream, respectively,

Fig. 3. The four expressions of the AR face database. (a) Neutral, (b) smile, (c) anger, and (d) scream.

TABLE 3
Facial Expression Results on the AR Database

	BAID	PMLT
smile	94%	**99%**
anger	86%	**98%**
scream	56%	**71%**

and, thus, fairly similar to the BAID results. This experiment supports the hypothesis.

Note that in this case our feature transformations are useless and might decrease the performance of the system. Indeed, the PMLT uses two types of transformations, grid and feature transformations, which "compete" to explain the observed variability and the performance of the PMLT could decrease if we allow feature transformations while no illumination variation is observed. We thus reran the PMLT by constraining the system to be in the feature state $\phi_{i,j} = [0 \cdots 0]^T$ at each position (i, j) and we did not observe any significant difference in performance. This shows than, even if no illumination variation is observed, the PMLT does not try to interpret facial expression variations as illumination variations.

6.4.4 Illumination

The robustness of the BAID and PMLT with respect to illumination variations was evaluated on the AR, PIE, and Yale B databases:

- For the AR database, sets 05, 06, and 07, which correspond, respectively, to the left light on, the right light on, and both lights on, were used as test data (see Fig. 4). The neutral expression was used as enrollment image.
- For the PIE database, experiments were carried out on the sets with and without ambient lighting, which will be later referred to as PIE 1 and PIE 2 (see Figs. 5 and 6). Only the images corresponding to the frontal camera were used. For each of the 68 persons, an image corresponding to the pure ambient lighting of PIE 1 was used as enrollment image and the 2×21 other conditions were used as test data.
- For Yale B, we also only used those images which correspond to the frontal camera. The image which corresponds to the flash which is directly in the optical axis of the camera was chosen as enrollment image and we used as test data 38 images which correspond to flashes which make an angle between 20 degrees and 77 degrees with the optical axis. The images of Yale B are very similar to the images of PIE 2.

Results are shown in Table 4. PMLT seems to almost always outperform BAID. Both algorithms perform very well on AR 05 and AR 06 and on PIE 1. We believe that the reason for the poor performance of both algorithms on AR 07 is the fact that, with both lights on, many images are overilluminated and have a very low contrast. The PIE 2 and Yale B, which both correspond to a flash when there is no background illumination, seem to be also fairly difficult

Fig. 4. The three illumination conditions of the AR face database. (a) Left light on. (b) Right light on. (c) Both lights on.

Fig. 5. Different illumination conditions for PIE 1, i.e., with ambient lighting.

Fig. 6. Different illumination conditions for PIE 2, i.e., without ambient lighting.

TABLE 4
Illumination Results on the FERET, PIE, and Yale B Databases

		BAID	PMLT
AR	05 + 06	94%	**100%**
	07	48%	54%
PIE	PIE 1	100%	100%
	PIE 2	54%	**65%**
Yale B		94%	93%

TABLE 5
Pose Results on the PIE Database

	BAID	PMLT
05 + 29	70%	**90%**
07 + 09	69%	**94%**
11 + 37	32%	**57%**

sets (remember that Yale B contains the face images of only 10 persons).

It is interesting to quantify the impact of the feature transformations on the performance in the presence of illumination variations. If we force the system to be at each position (i, j) in the state $\phi_{i,j} = [0 \cdots 0]^T$, then the performance of PMLT decreases significantly only for PIE 2 (54 percent) and Yale B (80 percent).

In the case of pure illumination variations, grid transformations are practically useless, in the same manner feature transformations were useless to model facial expressions. We thus ran the PMLT by constraining the system to be in the grid state $\tau_{i,j} = [0 \cdots 0]^T$ at each position (i, j) and did not observe any significant difference in performance. This shows than, even if no facial expression variation is observed, the PMLT does not try to compensate for illumination variations with grid transformations.

6.4.5 Pose

Although we did not train our system to be robust to pose variations, we think it is interesting to ascertain the robustness of the PMLT with respect to the pose as it is a source of large intraclass variability. Experiments were carried out on the PIE database. The test data consisted of the 68 persons. We chose the images with neutral expressions from 6 cameras: 05, 07, 09, 29, and 37. These test sets were grouped as follows: 07 and 09, 05 and 29, 11 and 37. These three sets correspond approximately to up or down rotations of the head of ± 15 degrees and to left or right rotations of ± 22 degrees and ± 45 degrees (see Fig. 7).

Results are presented in Table 5. The PMLT algorithm outperforms very significantly the BAID algorithm for all poses. Since we also suspected that the difference in performance was primarily due to the grid transformations of the PMLT, as was the case for facial expressions, we ran the rigid version of the PMLT and obtained respectively 72, 69, and 34 percent on 05 + 29, 07 + 09, and 11 + 37, respectively. The scores we obtain with the BAID and the rigid version of the PMLT are very comparable, thus validating our hypothesis.

It is interesting to see, that while the PMLT was trained only on the FA and FB sets of FERET which exhibit very little pose variability, it does manage to generalize on novel views. Note however that, even for the PMLT, the performance drops drastically for poses of approximately ± 45 degrees.

6.4.6 Pose and Illumination

Finally, we demonstrate the ability of our algorithm to deal with pose and illumination variations, and thus with grid and feature transformations, at the same time. Experiments were carried out on the Yale B face database. We used the data from the nine cameras for these experiments and images were divided into three sets according to the angle θ between the flash and the optical axis of the frontal camera: $20^o \leq \theta \leq 25^o$ for "illu 1," $35^o \leq \theta \leq 50^o$ for "illu 2," and $60^o \leq \theta \leq 77^o$ for "illu 3." Results are presented in Table 6.

The PMLT outperforms significantly the BAID under all other conditions, thus showing that the PMLT can deal with both grid and feature transformations at the same time.

7 Conclusion and Future Work

In this section, we first summarize the original approach introduced in this paper and our experimental findings. We will then consider two possible directions for future work.

7.1 Summary

In this article, we introduced a novel measure of "distance" between faces which involves the estimation of the set of possible transformations between face images of the same person. The global transformation is approximated with a set of local transformations under a constraint imposing consistency between neighboring local transformations. Local transformations and neighboring constraints are embedded within the probabilistic framework of a two-dimensional hidden Markov model. This general framework was specialized to the problem of face recognition and we focused on grid and feature transformations.

(a) (b) (c) (d) (e) (f)

Fig. 7. Different poses for PIE. (a) Camera 05. (b) Camera 07. (c) Camera 09. (d) Camera 11. (e) Camera 29. (f) Camera 37.

TABLE 6
Pose and Illumination Results on the Yale B Database

	pose	BAID	PMLT
Yale B	1	73%	**86%**
	2	60%	**70%**
	3	46%	**58%**

The performance of this probabilistic model of face mapping was assessed on a large data set consisting of four face databases (FERET, Yale B, PIE, and AR) and involving more than 10,000 images. A comparison was carried out with the Bayesian intra/extrapersonal classifier, which is one of the most successful approaches to face recognition to date, and it was shown that the proposed probabilistic model of face mapping compares favorably for facial expression, pose, and, to a lesser extent, illumination variations. More precisely, grid transformations are especially useful to model facial expressions and to deal with pose variations and feature transformations are useful to compensate for extreme illumination variations.

7.2 Clustering

We believe however that one of the limitations of our algorithm is its computational complexity. When running our nonoptimized code on a 2 Ghz Pentium 4 with 1 GB RAM, the comparison of two face images takes approximately 25 ms (once feature vectors are extracted from the template and query images). This is to be compared for instance to the computational cost of BAID for which the comparison of two images takes on the order of 0.1 ms. While it is possible to run our algorithm in the identification mode for a set of approximately 100 persons, for a larger set, the response time might be too long and, thus, incur an inconvenience for the user.

A possible solution to this problem is to perform clustering [1]. The basic idea is to group users into clusters and to perform the identification in two stages. When a new target image is added to the database, one computes the distance between this image and all cluster centroids and the image is associated to its nearest cluster. When a query image is probed, the first step consists in determining the nearest cluster and the second step involves the computation of the distances between the query image and the target images assigned to the corresponding cluster, thus reducing significantly the number of comparisons.

The clustering algorithm itself is performed offline in an unsupervised manner. Note that the clustering procedure, which involves 1) the computation of the distance between the training observations and the cluster centroids and 2) the reestimation of the cluster centroids, is heavily dependent on the chosen measure of distance. Until now our work has focused on the issue of distance computation between images and the "missing" stage to be able to perform clustering with the proposed distance is the centroid estimation. When using simple metrics such as the Euclidean distance, the centroid estimation consists in computing a simple average of the assigned observations. However, in the case of complex distances such as the distance induced by the probabilistic model of face mapping, computing the centroid is far from obvious. Note that, to alleviate the issue of cluster centroid estimation, one could make use of the concept of medoids [55], i.e., choose as a centroid of a given cluster one of the face images which is assigned to this cluster. While such an approach may give reasonable results, our preliminary work on the topic shows that, for the problem under consideration, an improved performance can be obtained when using centroids instead of medoids.

It should be underlined that the ability to compute the centroid of multiple face images has other potential applications than clustering. For instance, while enrollment data can be limited to one unique face image, if multiple images are available then the system should be able to deal with this additional information and to merge these images into a robust template. Also, as the face of a person generally varies slowly over time, except for radical punctual changes (due for instance to shaving), it could be of interest to perform a continuous unsupervised adaptation of the client models, i.e., if a user is accepted by the system with a high degree of confidence, then the test image could be used to update the template.

7.3 Applying PMLT to Other Domains

We believe that one of the strengths of the PMLT framework is its generality and that it has the potential to be extended to the retrieval of other types of images. Especially, within the field of biometrics, we envision to apply PMLT to the problem of automatic fingerprint recognition [56]. We now explain how to export PMLT to other problem domains and in particular to the case of fingerprint recognition. The two crucial issues are the choice of *feature vectors* and *local transformations*.

To choose a relevant set of features, it is necessary to understand what characterizes fingerprint images. A fingerprint is the pattern of ridges and furrows in the central region of the fingertip. Fingerprint recognition has been traditionally based on the extraction and matching of minute details (*minutiae*) associated with ridges and furrows [56]. However, such features are difficult to extract precisely and alternative representations of the fingerprint can be used. It has been shown that one could consider the pattern of ridges and furrows as a texture image and that Gabor features, which are particularly relevant for the analysis of textures, could be used to characterize fingerprint images [57].

Then, it is crucial to understand which variabilities have to be modeled and to choose our local transformations accordingly. Indeed, as outlined in Section 2, the measure of similarity primarily depends on the choice of local transformations. In the following, we will focus our attention on one type of variability: the elastic deformations of the fingerprint image incurred from the acquisition process. These deformations might change from one acquisition to another as they depend on the exact contact point but also on the pressure of the finger on the sensing device. As a first approximation, these distortions may be modeled with grid transformations in the same manner we modeled the elastic deformations incurred from expressions

for the problem of face recognition. However, we believe that such complex deformations would not be fully handled by grid transformations and that it might be useful to make use also of local rotation/scale transformations. Gabor features would be particularly appropriate to model such transformations. Indeed, each Gabor wavelet within the filter bank corresponds to an analysis of the image content in a given orientation and for a given scale. Thus, a small rotation or change in the scale could be readily interpreted as a shift of the energy between frequency bands. This phenomenon can be approximated by a linear transform (i.e., a matrix multiplication) of the Gabor feature vector. To summarize, elastic distortions of the fingerprint could be modeled accurately with a combination of grid transformations but also rotation and scale feature transformations.

ACKNOWLEDGMENTS

This work was supported in part by France Telecom Research and Development, by the US National Science Foundation under grant IIS-0329267, and by the University of California MICRO program, Applied Signal Technology, Inc., Dolby Laboratories, Inc., and Qualcomm, Inc. This work was completed while F. Perronnin was with the Multimedia Communications Department at Institut Eurecom.

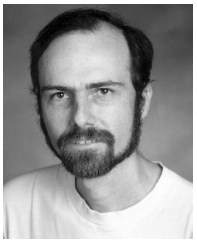
REFERENCES

[1] R. Duda, P. Hart, and D. Stork, *Pattern Classification,* second ed. John Wiley & Sons, 2000.
[2] A. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Trans. Circuits Systems Video Technology,* vol. 14, no. 1, Jan. 2004.
[3] R. Chellappa, C. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," *Proc IEEE,* vol. 83, no. 5, pp. 705-740, May 1995.
[4] J. Weng and D. Swets, *Biometrics: Personal Identification in Networked Society.* Kluwer, 1999.
[5] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The Feret Evaluation Methodology for Face Recognition Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 10, pp. 1090-1104, Oct. 2000.
[6] D. Blackburn, M. Bone, and P. Phillips, "Face Recognition Vendor Test 2000: Evaluation Report," technical report, 2001.
[7] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and M. Bone, "Face Recognition Vendor Test 2002: Evaluation Report," technical report, 2003.
[8] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition,* pp. 586-591, 1991.
[9] K. Etemad and R. Chellappa, "Face Recognition Using Discriminant Eigenvectors," *Proc. IEEE Int'l Conf. Acoustics Speech and Signal Processing,* vol. 4, pp. 2148-2151, 1996.
[10] D. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 831-836, Aug. 1996.
[11] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.
[12] J. Beveridge, K. She, B. Draper, and G. Givens, "A Nonparameteric Statistical Comparison of Principal Component and Linear Discriminant Subspaces for Face Recognition," *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition,* pp. 535-542, 2001.
[13] C. Liu and H. Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition," *IEEE Trans. Image Processing,* vol. 11, no. 4, pp. 467-476, Apr. 2002.
[14] R. Beveridge, D. Bolme, M. Teixeira, and B. Draper, *The CSU Face Identification Evaluation System Users' Guide Version 5.0,* technical report, Computer Science Dept., Colorado State Univ., May 2003.
[15] W. Zhao, "Robust Image Based 3d Face Recognition," PhD dissertation, Dept. of Electrical and Computer Eng., Univ. of Maryland, 1999.
[16] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 696-710, July 1997.
[17] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond Eigenfaces: Probabilistic Matching for Face Recognition," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 30-35, 1998.
[18] B. Moghaddam, "Principal Manfolds and Probabilistic Subspaces for Visual Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 6, pp. 780-788, June 2002.
[19] B. Moghaddam, C. Nastar, and A. Pentland, "A Bayesian Similarity Measure for Deformable Image Matching," *Image and Vision Computing,* vol. 19, pp. 235-244, 2001.
[20] M. Lades, J. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Würtz, and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Trans. Computers,* vol. 42, no. 3, pp. 300-311, Mar. 1993.
[21] M. Vissac, J.-L. Dugelay, and K. Rose, "A Novel Indexing Approach for Multimedia Image Databases," *Proc. IEEE Workshop Multimedia Signal Processing,* pp. 97-102, 1999.
[22] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression (Pie) Database," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* 2002.
[23] A. Georghiades, P. Belhumeur, and D. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 6 pp. 643-660, June 2001.
[24] A. Martínez and R. Benavente, "The AR Face Database," Technical Report 24, CVC, 1998.
[25] F.S. Samaria, "Face Recognition Using Hidden Markov Models," PhD dissertation, Univ. of Cambridge, Cambridge, U.K., 1994.
[26] A. Nefian, "A Hidden Markov Model-Based Approach for Face Detection and Recognition," PhD dissertation, Georgia Inst. of Technology, Atlanta, 1999.
[27] S. Eickeler, S. Müller, and G. Rigoll, "Recognition of Jpeg Compressed Face Images Based on Statistical Methods," *Image and Vision Computing,* vol. 18, no. 4, pp. 279-287, Mar. 2000.
[28] F. Cardinaux, C. Sanderson, and S. Bengio, "Face Verification Using Adapted Generative Models," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR),* pp. 825-830, 2004.
[29] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," *Proc. Int'l Conf. Spoken Language Processing,* vol. 2, pp. 1137-1140, 1996.
[30] A. Acero and X. Huang, "Speaker and Gender Normalization for Continuous-Density Hidden Markov Models," *Proc. IEEE Int'l Conf. Acoustics Speech and Signal Processing,* vol. 1, pp. 342-345, 1996.
[31] E. Bocchieri, "Phonetic Context Dependency Modeling by Transform," *Proc. Int'l Conf. Spoken Language Processing,* vol. 4, pp. 179-182, 2000.
[32] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications," *Proc. IEEE,* vol. 77, no. 2, pp. 257-286, Feb. 1989.
[33] S.-Z. Li, "Markov Random Field Models in Computer Vision," *Proc. IEEE European Conf. Computer Vision (ECCV),* vol. B, pp. 361-370, 1994.
[34] K. Abend, T. Harley, and L. Kanal, "Classification of Binary Random Patterns," *IEEE Trans. Information Theory (IT),* vol. 11, no. 4, pp. 538-544, Oct. 1965.
[35] S. Kuo and O. Agazzi, "Keyword Spotting in Poorly Printed Documents," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 8, pp. 842-848, Aug. 1994.
[36] J. Li, A. Najmi, and R. Gray, "Image Classification by a Two-Dimensional Hidden Markov Model," *IEEE Trans. Signal Processing,* vol. 48, no. 2, pp. 517-533, Feb. 2000.
[37] C. Miller, B. Hunt, M. Neifeld, and M. Marcellin, "Binary Image Reconstruction Via 2-D Viterbi Search," *Proc. IEEE Int'l Conf. Image Processing,* vol. 1, pp. 181-184, 1997.
[38] K. Hallouli, L. Likforman-Sulem, and M. Sigelle, "A Comparative Study Between Decision Fusion and Data Fusion in Markovian Printed Character Recognition," *Proc. IEEE Int'l Conf. Pattern Recognition (ICPR),* vol. 3, pp. 147-150, 2002.

[39] F. Perronnin, J.-L. Dugelay, and K. Rose, "Iterative Decoding of Two-Dimensional Hidden Markov Models," *Proc. IEEE Int'l Conf. Acoustics Speech and Signal Processing,* vol. 3, pp. 329-332, 2003.
[40] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the Em Algorithm," *J. the Royal Statistical Soc.,* vol. 39, no. 1, pp. 1-38, 1977.
[41] B. Horn, *Robot Vision.* McGraw-Hill, 1986.
[42] A.H.S.T. Kailath and B. Hassibi, *Linear Estimation.* Prentice Hall, 2000.
[43] F. Perronnin and J.-L. Dugelay, "From Turbo Hidden Markov Models to Turbo State-Space Models," *Proc. IEEE Int'l Conf. Acoustics Speech and Signal Processing,* 2004.
[44] R. Singh and B. Raj, "Tracking Noise via Dynamical Systems with a Continuum of States," *Proc. IEEE Int'l Conf. Acoustics Speech and Signal Processing (ICASSP),* vol. 1, pp. 396-399, 2003.
[45] J. Droppo and A. Acero, "Noise Robust Speech Recognition with a Switching Linear Dynamic Model," *Proc. IEEE Int'l Conf. Acoustics Speech and Signal Processing (ICASSP),* vol. 1, pp. 953-956, 2004.
[46] M. Grudin, "On Internal Representations in Face Recognition Systems," *Pattern Recognition,* vol. 33, no. 7, pp. 1161-1177, 2000.
[47] R. Gross, J. Shi, and J. Cohn, "Quo Vadis Face Recognition," *Proc. Workshop Empirical Evaluation Methods in Computer Vision,* 2001.
[48] Y. Adini, Y. Moses, and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 721-732, July 1997.
[49] M. Savvides and V. Kumar, "Illumination Normalization Using Logarithm Transforms for Face Authentication," *Proc. IAPR Audio- and Video-Based Biometric Person Authentication,* pp. 549-556, 2003.
[50] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Classifying Facial Actions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 10, pp. 974-989, Oct. 1999.
[51] B. Duc, S. Fischer, and J. Bigün, "Face Authentication with Gabor Information on Deformable Graphs," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 8, no. 4, pp. 504-516, Apr. 1999.
[52] HTK, Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk, Cambridge Univ., 2004.
[53] M. Teixeira and J.R. Beveridge, "An Implementation and Study of the Moghaddam and Pentland Intrapresonal/Extrapersonal Image Difference Face Recognition Algorithm," technical report, Colorado State Univ., 2003.
[54] X. Wang and X. Tang, "Unified Subspace Analysis for Face Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 679-686, 2003.
[55] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley and Sons, 1990.
[56] D. Maltoni, D. Maio, A.K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition.* Springer Verlag, 2003.
[57] S. Prabhakar, "Fingerprint Classification and Matching Using a Filterbank," PhD dissertation, Michigan State Univ., 2001.

Florent Perronnin received the engineering degree in 2000 from the Ecole Nationale Supérieure des Télécommunications, Paris, France. From 2000 to 2001, he was with the Panasonic Speech Technology Laboratory, Santa Barbara, California, first as an intern and then as a research engineer, working on speech and speaker recognition. He was then a PhD candidate within the Multimedia Communications Department of the Institut Eurécom, Sophia Antipolis, France, focusing his research on automatic face recognition. In 2003, he received a best student paper award at the IEEE International Conference on Image Processing, for the paper "Deformable Face Mapping for Person Identification." In 2004, he obtained his PhD degree from the Ecole Polytechnique Fédérale de Lausanne, Switzerland. He then joined the Xerox Research Centre Europe, Meylan, France, where he is currently a research engineer.

Jean-Luc Dugelay (M'94-SM'02) received the PhD degree in computer science in 1992 from the University of Rennes. His doctoral research was carried out, from 1989 to 1992, at the France Telecom Research Laboratory in Rennes (formerly CNET—CCETT). He then joined the Institut Eurécom (Sophia Antipolis), where he is currently a professor in the Department of Multimedia Communications. His research interests are in the area of multimedia signal processing and communications; including security imaging (i.e., watermarking and biometrics), image/video coding, facial image analysis, virtual imaging, face cloning, and talking heads. He is an author or coauthor of more than 65 publications that have appeared as journal papers or proceeding articles, three book chapters, and three international patents. He gave several tutorials on digital watermarking (coauthored with F. Petitcolas from Microsoft Research), biometrics (coauthored with J.-C. Junqua from Panasonic Research), and compression at major conferences. He has been an invited speaker and/or member of the program committee of several scientific conferences and workshops. He was technical cochair and organizer of the fourth workshop on Multimedia Signal Processing (Cannes, October 2001), and coorganizer of the workshop on Multimodal User Authentication (Santa Barbara, December 2003). His group is involved in several national and European projects related to biometrics. He is a senior member of the IEEE Signal Processing Society, an elected member of the IEEE SP IMDSP TC, and is currently an associate editor for the *IEEE Transactions on Multimedia*. He is a senior member of the IEEE.



Kenneth Rose (S'85-M'91-SM'01-F'03) received the PhD degree in 1991 from the California Institute of Technology. In 1991, he joined the Department of Electrical and Computer Engineering, University of California at Santa Barbara, where he is currently a professor. His main research activities are in information theory, source and channel coding, video and audio coding and networking, pattern recognition, and nonconvex optimization in general. He is also particularly interested in the relations between information theory, estimation theory, and statistical physics, and their potential impact on fundamental and practical problems in diverse disciplines. His optimization algorithms have been adopted and extended by others in numerous disciplines beside electrical engineering and computer science. He currently serves as an area editor for the *IEEE Transactions on Communications*, as well as a reviewing editor for source-channel coding. He cochaired the technical program committee of the 2001 IEEE Workshop on Multimedia Signal Processing. He was corecipient of the 1990 William R. Bennett Prize-Paper Award of the IEEE Communications Society, and of the 2004 IEEE Signal Processing Society Best Paper Award (in image and multidimensional signal processing). He is a fellow of the IEEE.